프레임 보간을 이용한 자기지도학습 기반 영상 주행기록계

Self-Supervised Visual Odometry via Frame Interpolation

이 세 빈[1] · 임 우 빈[2] · 윤 성 의[†]
Sebin Lee[1], Woobin Im[2], Sung-Eui Yoon[†]

Abstract: Self-supervised visual odometry (VO) has been proposed to learn a VO model without ground-truth. The photometric loss is a widely used loss function to train a VO model, which maximizes consistency between the original image and the reconstucted image by motion, and does not require VO ground truth. However, the photometric loss tends to suffer from a weak supervision signal, due to the lack of credible supervision sources to infer the relative movement. To address this issue, we propose a novel self-supervised VO utilizing frame interpolation, which takes advantage of t he frame interpolation technique to intensify the supervision signal of the photometric loss through fra me interpolation; we interpolate frames in the VO dataset, so that the frames are augmented with additional credible knowledge. We show the photometric loss with the additional supervision signal he lps both the pose and depth network trained more accurately. In experiments, we show that our method improves performance of visual odometry on KITTI odometry dataset in a reasonable gap.

Keywords: Visual Odometry, Depth Estimation, Self-Supervised Learning

1. 서 론

3차원 공간에서 로봇이 이동함에 있어, 로봇은 스스로의 위치를 6-DoF로 추정하는 능력이 요구된다. 이러한 능력을 위해 카메라의 영상을 사용하여 로봇 자신의 위치를 추정하는 영상 주행기록계(Visual Odoemetry, VO)가 컴퓨터 비전 및 로봇 분야에서 10년 넘게 활발히 연구되고 있다.

최근 딥러닝이 다양한 로봇 분야에 적용되면서, VO분야에서도 딥러닝을 활용한 연구들이 진행되고 있다. 대표적으로 지도학습으로 VO를 수행한 연구가 있지만, 대량의 참 값(Ground Truth) 데이터가 필요한 단점이 있다. 한편, 자기주도학습은 입력 데이터에서 학습에 필요한 학습 신호를 간접적으로 만들어 학습하므로 직접적인 참 값 데이터를 필요로 하지 않는 장점이 있어 최근엔 자기주도학습 방식의 VO가 주로 연구되고 있다.

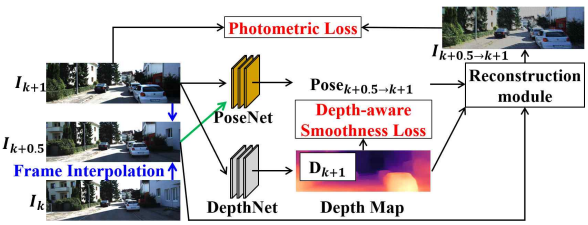
※ This project was funded by the MSIT under ITRC support program (IITP-2020-0-01460)
1. MS Student, Korea Advanced Institute of Science and Technology, Daejeon, Korea (seb.lee@kaist.ac.kr)
2. PhD Student, Korea Advanced Institute of Science and Technology, Daejeon, Korea (iwbn@kaist.ac.kr)
† Professor, Corresponding author: School of Computing, Korea Advanced Institute of Science and Technology, Daejeon, Korea (sungeui@kaist.edu)

기존의 자기주도학습 기반 VO는 약한 학습 신호 문제를 가지고 있는데, 그 원인 중 하나는 학습에 사용하는 손실함수가 갖는 정보가 적기 때문이다. 이를 해결하기 위해 본 논문에서는, 프레임 보간 기법으로 데이터를 증강하여, 학습 과정에 믿을 만한 정보를 추가 제공한다. 본 기법은 학습 신호를 강하게 만들고 결과적으로 영상 주행기록계의 성능을 향상시킨다.

2. 본 론

2.1 VO 심층신경망 학습

VO를 수행하는 심층신경망(Deep Neural Network, DNN) 모델 PoseNet은 주어진 입력 프레임 간의 카메라 6-DoF 상대 위치를 추정한다. 자기주도학습에 필요한 깊이 지도는 DNN 모델 DepthNet으로부터 추정한다. 위 두 모델은 Photometric Loss를 통해 훈련한다. 이상적으로, 두 프레임 I_s, I_t 간의 상대 정보(상대 위치, 깊이 지도)를 안다고 가정하면, I_s를 I_t로 복원한(Reconstruction) 이미지 $I_{s\to t}$는 I_t와 동일하다. Photometric Loss는 추정된 상대정보를 이용해 복원한 $I_{s\to t}$와 실제 이미지 I_t를 같도록 유도하고, PoseNet과 DepthNet을 함께 훈련시킨다.



[Fig. 1] Learning pipeline via frame interpolation. We utilize the interpolated image for better supervision signal.

움직이는 물체가 없는 정적 환경에서, PoseNet에 의한 상대 위치와 DepthNet에 의한 깊이 지도를 활용한 복원관계식[1]을 이용해 $I_{s\to t}$ 합성하고, 주 학습 신호인 Photometric Loss를 L1과 DSSIM의 가중합으로 나타내면 다음과 같다.

$$L_p = \sum_s (1-\alpha)\|I_t - I_{s\to t}\|_1 + \alpha\,\text{DSSIM}(I_s, I_{s\to t}) \qquad (1)$$

α는 가중치 비율을 나타낸다. 전체 손실함수 식은 상수 λ에 관한 가중합으로 아래와 같다.

$$L = L_p + \lambda L_s \qquad (2)$$

이때, L_s는, 깊이 지도에 대한 정규화(Regularization) 손실함수이며, Edge-aware Smoothness Loss[3]을 사용한다. VO는 정적 환경을 가정하기 때문에 Auto Mask[3]를 사용하여 움직이는 동적 물체에 대한 손실값을 L_p에서 제외한다.

2.2 프레임 보간을 이용한 학습

프레임 보간은 I_k와 I_{k+1} 를 입력하여 중간 프레임 $I_{k+0.5}$을 생성하는 과제이다. 이 과제는 자기지도학습을 통해 학습되며, 학습된 모델은 프레임 간의 상대적인 정보를 간접적으로 내재하고 있기 때문에 다른 과제의 모델에 유용한 정보를 줄 수 있다.

프레임 보간을 이용한 학습 방법의 파이프라인은 [Fig. 1]과 같다. 프레임 보간으로 생성한 중간 프레임을 Source 이미지 I_s, $s \in \{k-1, k-0.5, k+0.5, k+1\}$에 포함시켜, Target 이미지 $I_t = I_k$에 대해서 식 (2)의 손실함수를 적용한다. 이를 통해 프레임 보간 모델의 지식(Knowledge)이 VO 모델로 전달된다. 보간된 프레임을 추가적으로 고려하는 것은 신뢰할 수 있는 학습 신호를 강화시켜 학습된 모델이 더 나은 Local Minimum으로 수렴시킨다. 또한 이를 데이터 증강(Data Augmentation)으로 볼 수 있는데, 결과적으로 VO모델이 더 많은 데이터를 학습할 수 있어 성능을 향상시킨다. 게다가 제안한 방법은 기존 자기지도학습 기술에 접목하기 용이하므로 확장성에 있어 장점이 있다.

3. 실험 및 결과

기존의 다른 논문[1]과 마찬가지로 DepthNet과 PoseNet을 KITTI odometry dataset의 00-08을 이용해 학습하였다. 식 (1)의 α는 0.85, 식 (2)의 λ는 0.1, Optimizer는 Adam에 $\beta_1 = 0.9$, $\beta_2 = 0.999$, 학습률 0.0001을 적용하였으며, 200 에폭 동안 전체 손실함수인 식 (2)를 이용해 학습시켰다. 프레임 보간 모델은 KITTI raw dataset으로 자기지도학습을 시킨 뒤 모델 파라미터를 동결시켰다. 기존 논문[1]과 마찬가지로 짧은 이미지 시퀀스에 대한 ATE(Absolute Trajectory Error)를 KITTI odometry dataset의 09-10을 이용해 측정하였으며, 결과는 아래의 [Table 1]과 같다. 본 연구에서 제안한 방법은 ATE를 기준으로 SfMLearner[1] 보다 약 50%, GeoNet[2]보다 약 16% 가량 낮은 오차를 보여주고 있으며, VO성능 향상을 확인할 수 있다.

[Table 1] Absolute Trajectory Error for short snippets

Method	Sequence 09	Sequence 10
SfMLearner[1]	0.021±0.017	0.020±0.015
GeoNet[2]	0.012±0.007	0.012±0.009
Ours	**0.010±0.008**	**0.010±0.006**

4. 결 론

본 논문은 프레임 보간을 이용하여 자기지도학습 기반 VO 알고리즘을 제안하였다. 프레임 보간된 이미지를 통해 추가적인 학습 신호를 DNN에 전달함과 동시에 데이터 증강 효과를 부가적으로 얻어 성능을 향상시켰다. 향후 프레임 보간 기술을 최적화하여, 보간된 데이터의 높은 신뢰도를 바탕으로 VO 성능을 향상시키는 연구를 진행하려 한다.

References

[1] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised Learning of Depth and Ego-Motion From Video," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. *661 2-6619*, 2017.

[2] Z. Yin and J. Shi, "GeoNet: Unsupervised Learning of Dense Depth, Optical Flow and Camera Pose," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, pp. *1983-1992*, 2018.

[3] C. Godard, O. M. Aodha, M. Firman, G. Brostow, "Digging Into Self-Supervised Monocular Depth Estimation," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), pp. *3827-3837*, 2019.